Prior distributions for variance-covariance matrices in hierarchical models

Edoardo Costantini

April 26, 2019

1 Introduction

Fully-Bayesian analysis has grown in popularity over the last few decades (look for reference). In particular, fully Bayesian approaches to hierarchical models are gaining momentum thanks to the wider accessibility of computing power (ehh.. verify!). The field is flourishing with both in terms of practical applications and theoretical interest.

Among the most active theoretical discussions, there is a long lasting debate on the value of subjective versus objective Bayesian perspectives (see Berger, 2006). While the former approach is often defended as the more truly "Bayesian", allowing the analyst to include prior information in the estimation process, the latter is appropriate when a researchers wants to express lack of subjective information on the model parameters of interest. The present work is not aiming to contribute to this debate directly, but rather is meant to focus on a particular issue pertaining to the objective framework.

When subscribing to the "objective" framework, a researcher can reflect the lack of prior information, relating to the parameters, by defining prior distributions that are meant to be minimally informative in some sense. Many theoretical contributions have tried to indicate what distributions, in which situations, are more apt to achieve such a goal. These endeavors have produced practical guidelines for researchers to assist them in the crucial decision prior distributions represent in Bayesian analysis.

In making this decision, two concepts are fundamental: conditional conjugacy and informativeness of a prior distribution. A family of priors for a parameter is conditionally conjugate when the conditional posterior of said parameter is also in that family of distributions. An uninformative prior is a reference prior distribution defined in such a way that the posterior inference is not influence in any way by it. Related to this concept, is that of a weakly informative prior. This is a proper distribution (it integrates to 1) but is set up so that "the information it does provide is intentionally weaker than whatever actual prior knowledge is available" (Gelman, 2006). This def-

inition is rooted in the convenience, one may even say need, for a reference posterior distribution, by which is meant a distribution, obtained by employing an improper prior, that approximates a posterior that would have been obtained if a proper prior, describing initial vague information, had been used (Bernardo, 1979). In this paper, we will focus on proper priors, but the principle of uninformativeness still applies: objective Bayesian analysis needs a posterior distribution that does not incorporate the researcher's personal beliefs so that it is possible to asses the relevance of initial information.

Much attention has been dedicated in the literature to propose prior distributions for scale parameters that are alternatives to the overused inverse Gamma distribution (reference dump here). This effort finds its *raison d'être* in the undesirable shrinkage of the posterior distributions towards zero, even when the parameters of an inverse Gamma prior are specified to achieve weak informativeness (Brown and Draper, 2006; Gelman, 2006). Many alternatives have been proposed but the focus has always been on hierarchical models with single random effects, even when the declared intent was to speak of cases of higher dimensionality (Kass and Natarajan, 2006)

My work contributes to this literature by exploring the degree of uninformativity achieved by different prior distributions for the covariance matrix of a vector of random effects in a hierarchical model.

2 The hierarchical model of interest

I worked with a two-level normal model of repeated observations y_{ij} with individual (clusters) effects $\boldsymbol{b}_i$

$$\begin{aligned} y_{ij} &= \boldsymbol{x}_{ij}^T\boldsymbol{\theta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i + \epsilon_{ij} \\ \boldsymbol{b}_i &\sim N(\mathbf{0}, \boldsymbol{\Psi}) \\ \epsilon_{ij} &\sim N(0, \sigma^2) \end{aligned} \tag{1}$$

with $\boldsymbol{b}_i$ a vector of ranodm effects, and $\boldsymbol{\Psi}$, and σ^2 representing the model hyperparameters. The model is fitted to a repeated observations data set where the main predictor is *time* (x_1) and there is a dichotomous covariate x_2 that interacts with it. θ is a vector of fixed effects containing fixed intercept, time, covariate, and interaction effects. This model is a random intercept and random slope model, with $\boldsymbol{b}_i = [b_{i0}, b_{i1}]'$ a vector containing the cluster (individual) specific effects.

3 Weakly informative prior distributions in hierarchical models

The weak informativeness of a prior is inherently a provisional concept. It makes sense to fit a model with a specific prior definition, judge whether the the posterior distribution makes sense (e.g. the values allowed by the posterior should be in a plausible range for the parameter of interest, and the posterior itself should not be shrieked to zero by the prior); and adjust the prior specification if the posterior does not make sense. In other words, if the posterior obtained using a given prior spreads over an unrealistic range of values for the parameter (e.g. excessively large variances), or if it is shrieked towards an implausible small number, then it is reasonable to reconsider the prior choice.

In the case of scalar random effects, such as for a random intercept model (as opposed to the vector of random effects that characterizes random intercept ad slope models), it has been previously shown that the Inverse-gamma distribution with parameters $\alpha = \epsilon$, $\beta = \epsilon$, commonly considered non informative, does not look to uphold to such a property (Gelman, 2006).

A fundamental issue of Bayesian Statistics is the definition uninformative priors. It is common to distinguish between objective and subjective Bayesian approaches to the definition of priors (elaborate more). As Bayesian approaches to data analysis become more and more widespread a need for standard ways of defining priors for analyses becomes more and more relevant. While this issue might be easier for some types of analysis (elaborate more), when models become more complicated the issue becomes non-trivial. In particular there is no standard way of choosing an uninformative prior for the random effects variance-covariance matrix parameter in mixed effects models.

Publications have flourished on the matter but have often focused on the simpler case of a scalar random effect, the random intercept model. Here I will explain Brown and Draper 2006, Gelman 2006, Kass and Natarajan 2006 solutions. Some publications extended the reasoning to the vector random effects case, elaborating on the definition of a prior distribution for the variance covariance matrix (henceforth referred to as $\boldsymbol{\Psi}$), as random intercept and random slopes models.

Here I define the Bayesian model I want to use to test the performance of the different priors. The features I want to include are the following: continuous outcome, any number/measurement scale.

Let's start with the **model**

$$\begin{aligned} y_{ij} &= \boldsymbol{x}_{ij}^T\boldsymbol{\theta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i + \epsilon_{ij} \\ \boldsymbol{b}_i &\sim N(\mathbf{0}, \boldsymbol{\Psi}) \\ \epsilon_{ij} &\sim N(0, \sigma^2) \end{aligned} \tag{2}$$

(Vectors are in bold, matrix are capital Greek letters).

I'm going to define the following **priors**:

$$p(\boldsymbol{\theta}) \propto 1 \tag{3}$$

$$p(\sigma^2) \propto \sigma^{-2} \tag{4}$$

For what concerns the random effects variance covariance matrix, different priors are tested. In particular we used:

inverse-Wishart prior

$$p(\boldsymbol{\Psi}) \propto IW(\nu, S_0) \tag{5}$$

where $\boldsymbol{\Psi}$ is the matrix of variance-covariance hyperparameters defined in model 1; S_0 is a $k \times k$ matrix, usually considered as a prior guess for the covariance matrix. The diagonal elements of $\boldsymbol{\Psi}$ are inverse-Wisahrt distributed as well. In particular, the $k_1 \times k_1$ upper-left triangular sub-matrix $\boldsymbol{\Psi}_{11}$ is $IW(\nu-(k-k_1), S_{011})$ distributed. Furthermore, note that in our case the diagonal elements are the random intercept and slope parameters, and therefore are scalar. For a univariate case, the inverse Wishart distribution simplifies to an inverse Gamma with parameters $\alpha = \frac{\nu}{2}, \beta = \frac{S_{0kk}}{2}$. This means that $\sigma^2_{i0} \sim IW(\nu - 1, S_{011})$ becomes $\sigma^2_{i0} \sim IG(\frac{\nu-1}{2}, \frac{S_{0ii}}{2})$. By choosing $\nu = k - 1 + \epsilon$ and $S_{011} = k - 1 - 1 + \epsilon$ we obtain a prior on the variance parameters equivalent to the one $IG(\epsilon, \epsilon)$ studied by Gelman (2006). Note that k being equal to two (there are two random effects) in model 1, the inverse Wishart prior for $\boldsymbol{\Psi}$, that grants an $IG(\epsilon, \epsilon)$ on the variance parameters, is $IW(1 + \epsilon, \epsilon I_2)$.

- inverse-Wishart

$$p(\boldsymbol{\Psi}) \propto IW(\nu, S_0) \tag{6}$$

where we choose $\nu = k - 1 + e$, and $S_0 = diag(k - 1 + e)$, following indications by Gelman *et al* (2014). Given a $\boldsymbol{\Psi} \sim IW(\nu, S_0)$, where $\boldsymbol{\Psi}$ and S_0 are $k \times k$ matrices, it is known that $\boldsymbol{\Psi}_{11}$, the $k_1 \times k_1$ upper-left triangular sub-matrix of $\boldsymbol{\Psi}$, has an inverse-Wishart distribution as well. In particular, $\boldsymbol{\Psi}_{11} \sim IW(\nu - (k - k_1), S_{011})$. Furthermore, for a univariate case ($k = 1$), we know that an inverse Wishart distribution simplifies to an inverse Gamma with parameters $\alpha = \frac{\nu}{2}, \beta = \frac{S_{0kk}}{2}$. With the goal of resembling as close as possible what Gelman 2006 did, we try to define inverse wishart priors for $\boldsymbol{\Psi}$, such that the the marginal disitbrution is as close as possible to the IG(e, e) used by Gelman. This goal is achieved by setting $\nu = k - 1 + e$ and $S_{011} = k - 1 + e$, which makes the marginal distribution on the variance components (diagonals of $\boldsymbol{\Psi}$) $\sigma^2_{ii} \sim IG(\frac{\nu-1}{2}, \frac{S_{0ii}}{2})$. The inverse-Wishart priors we defined are:

Prior Description	ν	S_0
1. IW educated	2	educated guess
2. IW uninformative	$k-1+e$	$(k-1+e-1)\times$ diag(k)

1, .01, and .001 are then used as values of e.

- inverse-Wishart *a là* Huang and Wand

$$\begin{aligned} p(\boldsymbol{\Psi}|a_1, a_2) &\propto IW(\nu + k - 1, 2\nu \times diag(1/a_1, 1/a_2)), \\ a_k &\propto IG(1/2, 1/A_k^2), \end{aligned} \tag{7}$$

with $\nu = 2$ and $\boldsymbol{A} = [1000, 1000]$. The marginal distribution of any standard deviation term in $\boldsymbol{\Psi}$ is Half-$t(\nu, A_k)$ and, when choosing $\nu = 2$, the marginal distribution on the correlation term is uniform on (-1, 1), see property 2 to 4 in Huang and Wand (2013, p. 442). Furthermore, according to Huang and Wand (2013, p. 441) arbitrarily large values for a_k lead to arbitrarily weak priors on the standard deviation term. Hence, our choices for the parameters of this prior are:

Prior Description	ν	$\boldsymbol{A}$
3. IW a là HW	2	$[1000, 1000]$

- Matrix-F variate

Following Mulder, Pericchi 2018, I defined a matrix-F variate distribution as a prior for the covariance matrix of the random effects $\boldsymbol{\Psi}$

$$\begin{aligned} p(\boldsymbol{\Psi}) &\propto F(\boldsymbol{\Psi}; \nu, \delta, \boldsymbol{B}) \\ &\propto \int IW(\boldsymbol{\Psi}; \delta + k - 1, \Sigma) \times W(\boldsymbol{\Sigma}; \nu, \boldsymbol{B}) d\boldsymbol{\Sigma} \end{aligned} \tag{8}$$

with degrees of freedom $\nu > k - 1$, $\delta > 0$, and $\boldsymbol{B}$ a positive definite scale matrix that functions as prior guess. Different strategies can be followed in trying to achieve vagueness of this prior. In the literature, the improper prior $(\sigma^2)^{-\frac{1}{2}}$ has been proposed for the random effects variance (Berger, 2006; Berger and Strawderman, 1996). Placing a matrix-F prior on $\boldsymbol{\Psi}$, one can approximate the improper $|\boldsymbol{\Psi}|^{-\frac{1}{2}}$ by choosing $\nu = k$, the smallest allowed integer, $\boldsymbol{B} = b\boldsymbol{I}_k$ and letting $b \to +\infty$ for fixed values of δ.

Another approach might be trying to achieve a flat prior on the standard deviations of the random effects $p(\sigma) \propto 1$, through some other proper neighbor definition. Considering $\boldsymbol{\Psi} \sim matrix{-}F(\nu, \delta, \boldsymbol{B})$, it is known that the marginal distribution of diagonal elements of $\boldsymbol{\Psi}$,

the variance components σ^2_{jj}, is univariate $F(\nu, \delta, b_{jj})$ which is equivalent to $p(\sigma_{jj}; \nu, \delta, b) \propto \sigma^{\nu-1}(1+\sigma^2/b)^{-\frac{\nu+\delta}{2}}$. By choosing $\nu = k-1+\epsilon$, with say $\epsilon = .001$, $\delta = 1$, and $b \to \infty$, $p(\sigma) \propto 1$ is obtained. Hence, the specification $\boldsymbol{\Psi} \sim matrix - F(\nu = 1.001, \delta = 1, \boldsymbol{B} = 1e3\boldsymbol{I}_2)$ grants a proper prior neighbor to a flat prior on the standard deviation of the random effects.

It is also interesting to approach the vagueness issue by defining a vague prior weakly centered around an educated guess. Such educated guess can be obtained by plotting fitted regression lines to all the clusters in the data set that is under analysis, and trying to visually assess the intercepts and the slope variance. For example, consider the data set on depression that will be used to present the first empirical results. *here add plot and describe how you arrived to the educated guess* We can then try to achieve vagueness of the prior distribution by defining ν as small as possible. Hence, another prior definition that was used in this paper, set $\nu = k-1+\epsilon, \delta = 1$ and $\boldsymbol{B} = matrix(21, 0, 0, 9)$.

Finally, we could also specify a prior that uses the empirical guess defined by Kass and Natarajan (2006). To do so, we specify the vague matrix-F prior with scale matrix $\boldsymbol{R}^*$, and $\nu = 2$.

The following table summarizes the matrix-F specifications that grant some form of non informativeness and that are compared in this paper.

Prior Description	ν	δ	S_0
$\lvert\boldsymbol{\Psi}\rvert^{-\frac{1}{2}}$	2	1	$10^3 \times \boldsymbol{I}_2$
$p(\sigma) \propto 1$	1.001	1	$10^3 \times \boldsymbol{I}_2$
vaguely centered around educated guess	$k-1+\varepsilon$	ε	educated guess
vaguely centered around empirical R^*	2	1	$\boldsymbol{R}^*$

ε is set to each 1, and .1

4 Results

A good performance is shows by posteriors that give a reasonable range for the values of the parameters without having extremely long right-tails and without having excessively high peaks at zero.

In the figures, the empty circle represents the REML estimates of the parameter of interest. REML is used becasue it's the equivalent procedure to RIGLS, one of the most important ML estimation methods for Generalized Linear Models. RIGLS coincide with REML for all Gaussian models such as the one described in this article (Goldstein, 2010; Browne Draper, 2006).

When discussing the proper neighbor prior in the matrix F case for the SD of the variance, you should highlight that it is not an ideal situation, the role of a weakly informative prior is to regularize the posterior distribution so as to keep it into reasonable bounds (Gelman et al 2014

book). As a general guideline, priors should be made more precises as posteriors are more vague (which happens when fewer data are available).

When discussing the IW prior with e you can highlight that, compared to the uninformative matrix F priors, the range of values supported by the posterior distribution is decidedly more concentrated below 2 with a sharp peak towards 0.

In comparison, even when data is scarce (last few conditions), the posterior distributions obtained with the matrix F distribution allow for plausible values in a larger range that is more meaningful, while regularized by the prior to avoid impossible high values (which does happen when using the proper neighbor prior).

When discussing the priors for the correlation in the IW and matrix-F case, you should point out that the beta priors obtained, are weakly informative according to the symmetry principle: symmetrical prior distributions do not pull the posteriors in any particular direction.

Derivation of Full Conditional Posterior Distribution

The derivation of the conditional posterior follows.

Full conditional for $\boldsymbol{\theta}$ (fixed effects)

Let's start with

$$p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{b}_i, \boldsymbol{\Psi}, \sigma^2) = p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{b}_i, \boldsymbol{\Psi}, \sigma^2)p(\boldsymbol{\theta})$$

where

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Psi}, \sigma^2) &= \prod_{i=1}^{n}\prod_{j=1}^{J} p(y_{ij}|\boldsymbol{\theta}^T\boldsymbol{x}_{ij} + \boldsymbol{b}_i^T\boldsymbol{z}_{ij}, \boldsymbol{\Psi}, \sigma^2) \\ &\propto exp(-\frac{1}{2\sigma^2}SSR) \end{aligned}$$

and

$$SSR = \sum_{i=1}^{n}\sum_{j=1}^{J}(y_{ij} - \boldsymbol{\theta}^T\boldsymbol{x}_{ij} - \boldsymbol{b}_i^T\boldsymbol{z}_{ij})^2$$

where can rewrite y_{ij} as $\tilde{y}_{ij}$, with $\tilde{y}_{ij} = y_{ij} - \boldsymbol{b}_i^T\boldsymbol{z}_{ij}$ which makes SSR:

$$\begin{aligned} SSR &= \sum_{i=1}^{n}\sum_{j=1}^{J}(\tilde{y}_{ij} - \boldsymbol{\theta}^T\boldsymbol{x}_{ij})^2 \\ &= (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\theta})^T(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\theta}) \\ &= \tilde{\boldsymbol{y}}^T\tilde{\boldsymbol{y}} - 2\boldsymbol{\theta}^T\boldsymbol{X}\tilde{\boldsymbol{y}} + \boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta} \end{aligned}$$

Hence,

$$p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Psi}, \sigma^2) \propto exp(-\frac{1}{2\sigma^2}[-2\boldsymbol{\theta}^T\boldsymbol{X}\tilde{\boldsymbol{y}} + \boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta}])$$

Combining this with the prior we obtain:

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\Psi},\sigma^2) &\propto exp(-\frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta}+\boldsymbol{\theta}^T\boldsymbol{X}\tilde{\boldsymbol{y}}) \\ \boldsymbol{\theta}|. &\sim \boldsymbol{N}\left(\frac{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}\tilde{\boldsymbol{y}}}{\sigma^2},\frac{(\boldsymbol{X}^T\boldsymbol{X})^{-1}}{\sigma^2}\right) \end{aligned} \tag{9}$$

Full conditional for $\boldsymbol{b}_i$ (random effects)

To derive this one we can start from:

$$p(\boldsymbol{b}_i|\boldsymbol{y},\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\Psi},\sigma^2) = p(\boldsymbol{y_i}|\boldsymbol{\theta},\boldsymbol{b}_i,\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\Psi},\sigma^2)p(\boldsymbol{b}_i)$$

We know that

$$p(\boldsymbol{y}_i|.) = \prod_{j=1}^{J} p(y_{ij}|\boldsymbol{\theta}^T\boldsymbol{x}_{ij}+\boldsymbol{b}_i^T\boldsymbol{z}_{ij},\sigma^2) \propto exp(-\frac{1}{2\sigma^2}SSR_i)$$

with

$$SSR = \sum_{j=1}^{J}(y_{ij}-\boldsymbol{\theta}^T\boldsymbol{x}_{ij}-\boldsymbol{b}_i^T\boldsymbol{z}_{ij})^2$$

and we can rewrite y_{ij} as $\tilde{y}_{ij} = y_{ij} - \boldsymbol{\theta}^T\boldsymbol{x}_{ij}$, which would make SSR be

$$\begin{aligned} SSR &= \sum_{j=1}^{J}(\tilde{y}_j-\boldsymbol{\theta}^T\boldsymbol{x}_j)^2 \\ &= (\tilde{\boldsymbol{y}}-\boldsymbol{b}_i^T\boldsymbol{Z}_i)^T(\tilde{\boldsymbol{y}}-\boldsymbol{b}_i^T\boldsymbol{Z}_i) \\ &= \tilde{\boldsymbol{y}}^T\tilde{\boldsymbol{y}} - 2\boldsymbol{b}_i^T\boldsymbol{Z}_j\tilde{\boldsymbol{y}} + \boldsymbol{b}_i^T\boldsymbol{Z}_i^T\boldsymbol{Z}_j\boldsymbol{b}_i \end{aligned}$$

Hence,

$$p(\boldsymbol{y}_i|.) \propto exp(-\frac{1}{2\sigma^2}[-2\boldsymbol{b}_i^T\boldsymbol{Z}_j\tilde{\boldsymbol{y}}+\boldsymbol{b}_i^T\boldsymbol{Z}_i^T\boldsymbol{Z}_j\boldsymbol{b}_i])$$

We also know that in this case, the "prior" is

$$p(\boldsymbol{b}_i) \propto N(\boldsymbol{0},\boldsymbol{\Psi}) \propto exp(-\frac{1}{2}[-2\boldsymbol{b}_i^T\boldsymbol{\Psi}^{-1}\boldsymbol{0}+\boldsymbol{b}_i^T\boldsymbol{\Psi}^{-1}\boldsymbol{b}_i])$$

In conclusion, combining the sampling model and the prior, we get:

$$\begin{aligned} p(\boldsymbol{b}_i|.) &\propto exp(-\frac{1}{2\sigma^2}[-2\boldsymbol{b}_i^T\boldsymbol{Z}_j\tilde{\boldsymbol{y}}+\boldsymbol{b}_i^T\boldsymbol{Z}_i^T\boldsymbol{Z}_j\boldsymbol{b}_i]-\frac{1}{2\sigma^2}[-2\boldsymbol{b}_i^T\boldsymbol{Z}_j\tilde{\boldsymbol{y}}+\boldsymbol{b}_i^T\boldsymbol{Z}_i^T\boldsymbol{Z}_j\boldsymbol{b}_i]) \\ \boldsymbol{b}_i|. &\propto \boldsymbol{N}\left(\left(\Psi^{-1}+\frac{\boldsymbol{Z}_i^T\boldsymbol{Z}_i}{\sigma^2}\right)^{-1}\left(\boldsymbol{\Psi}^{-1}\boldsymbol{0}+\frac{\boldsymbol{Z}_i^T\tilde{y}_i}{\sigma^2}\right),\left(\Psi^{-1}+\frac{\boldsymbol{Z}_i^T\boldsymbol{Z}_i}{\sigma^2}\right)^{-1}\right) \end{aligned} \tag{10}$$

Full conditional for σ^2 (error variance)

The full conditional posterior can be expressed as:

$$p(\sigma^2|\boldsymbol{y},\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{b}_i,\boldsymbol{\Psi}) = p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{b}_i,\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\Psi},\sigma^2)p(\sigma^2)$$

The sampling model is the same we saw for the full conditional distribution of $\boldsymbol{\theta}$:

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Psi}, \sigma^2) &= \prod_{i=1}^{n} \prod_{j=1}^{J} p(y_{ij}|\boldsymbol{\theta}^T \boldsymbol{x}_{ij} + \boldsymbol{b}_i^T \boldsymbol{z}_{ij}, \boldsymbol{\Psi}, \sigma^2) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{J} (2\pi\sigma^{-2})^{-\frac{1}{2}} exp(-\frac{(y_{ij} - \boldsymbol{\theta}^T \boldsymbol{x}_{ij} - \boldsymbol{b}_i^T \boldsymbol{z}_{ij})^2}{2\sigma^2}) \end{aligned}$$

However, we are now interested in σ^2, hence

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Psi}, \sigma^2) &\propto (\sigma^2)^{-\frac{N}{2}} exp(-\frac{\sum_{i=1}^{n} \sum_{j=1}^{J} (y_{ij} - \boldsymbol{\theta}^T \boldsymbol{x}_{ij} - \boldsymbol{b}_i^T \boldsymbol{z}_{ij})^2}{2\sigma^2}) \\ &\propto (\sigma^2)^{-\frac{N}{2}} exp(-\frac{1}{2\sigma^2} SSR) \end{aligned}$$

where $N = \sum_i^n nj_i$ is the entire sample size (all observations within all clusters). The prior for σ is given above, and therefore we can write the full conditional posterior as:

$$\begin{aligned} p(\sigma^2|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{b}_i, \boldsymbol{\Psi}) &\propto (\sigma^2)^{-\frac{N}{2}-1} exp(-\frac{1}{2\sigma^2} SSR) \\ \sigma^2|. &\sim IG(\frac{N}{2}, \frac{SSR}{2}) \end{aligned} \tag{11}$$

Full conditional for $\boldsymbol{\Psi}$ (random effects variance covariance matrix)

Here, we need to write down the posteriors for the different priors we specified. First, let us define the sampling model for the random effects.

$$\begin{aligned} \begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} &= \boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{\Psi}) \\ p(\boldsymbol{b}_1, \boldsymbol{b}_2|\boldsymbol{\Psi}) &\propto |\boldsymbol{\Psi}|^{-\frac{n}{2}} exp\left(-\frac{1}{2} tr(\boldsymbol{S}_b \boldsymbol{\Psi}^{-1})\right) \end{aligned} \tag{12}$$

where $\boldsymbol{S}_b$ is $\Sigma_i \boldsymbol{b}_i \boldsymbol{b}_i^T$

- given the inverse-Wishart prior

$$\begin{aligned} p(\boldsymbol{\Psi}) &\propto IW(\nu, \boldsymbol{S}_0) \\ &\propto |\boldsymbol{\Psi}|^{-\frac{(\nu+k+1)}{2}} exp\left(-\frac{1}{2} tr(\boldsymbol{S}_0 \boldsymbol{\Psi}^{-1})\right) \end{aligned}$$

the full conditional posterior of $\boldsymbol{\Psi}$ is

$$\begin{aligned} p(\boldsymbol{\Psi}|.) &\propto |\boldsymbol{\Psi}|^{-\frac{(\nu+n+k+1)}{2}} exp\left(-\frac{1}{2} tr([\boldsymbol{S}_0 + \boldsymbol{S}_b]\boldsymbol{\Psi}^{-1})\right) \\ &\propto IW(\nu + n, \boldsymbol{S}_0 + \boldsymbol{S}_b) \end{aligned} \tag{13}$$

where $\nu = 2$

- inverse-Wishart *a là* Huang and Wand

$$
\begin{aligned}
p(\boldsymbol{\Psi}|a_1, a_2) &\propto IW(\nu + k - 1, 2\nu diag(1/a_1, 1/a_2)), \\
a_k &\propto IG(\eta, 1/A_k^2) \\
p(\boldsymbol{\Psi}) &\propto |\boldsymbol{\Psi}|^{-\frac{(\nu+k-1+1)}{2}} exp\left(-\frac{1}{2} tr(2\nu diag(1/a_1, 1/a_2)\boldsymbol{\Psi}^{-1})\right) \\
&\times \left(\frac{1}{a_1}\right)^{\eta+1} exp\left(-\frac{1}{A_1^2 a_1}\right) \times \left(\frac{1}{a_2}\right)^{\eta+1} exp\left(-\frac{1}{A_2^2 a_2}\right)
\end{aligned}
$$

the full conditional posterior of $\boldsymbol{\Psi}$ is

$$
\begin{aligned}
p(\boldsymbol{\Psi}|.) &\propto |\boldsymbol{\Psi}|^{-\frac{(\nu+k-1+n+1)}{2}} exp\left(-\frac{1}{2} tr([\boldsymbol{S}_b + 2\nu diag(1/a_1, 1/a_2)]\boldsymbol{\Psi}^{-1})\right) \\
&\propto IW(\nu + k - 1 + n, \boldsymbol{S}_b + 2\nu diag(1/a_1, 1/a_2)) \qquad (14) \\
p(a_k|.) &\propto IG\left(\eta(\nu + k), \nu\left(\boldsymbol{\Psi}_{kk}^{-1} + \frac{1}{A_k^2}\right)\right)
\end{aligned}
$$

where $\eta = \frac{1}{2}, \nu = 2, k = 2$, and n is the number of clusters (individuals). (For the conditional posterior of a_k refer to Huang and Wand (2013), section 4.2).

- Matrix-F variate

  Following section 2.3 in Mulder and Pericchi (2018), instead of working directly with the $\boldsymbol{\Psi} \sim F(\nu, \delta, \boldsymbol{B})$ we apply the parameter expansion defined above (see section on priors) and model it as $\boldsymbol{\Psi} \sim IW(\delta + k - 1, \boldsymbol{\Omega})$ with $\boldsymbol{\Omega} \sim W(\nu, \boldsymbol{B})$. With this parameter expansion, the conditional priors are:

$$
\begin{aligned}
\boldsymbol{\Psi}|\boldsymbol{\Omega} &\sim IW(\delta + k - 1, \boldsymbol{\Omega}) \\
\boldsymbol{\Omega}|\boldsymbol{\Psi} &\sim W(\nu + \delta + k - 1, (\boldsymbol{\Psi}^{-1} + \boldsymbol{B}^{-1})^{-1})
\end{aligned}
$$

  which makes the full conditional posterior of:

$$
\begin{aligned}
\boldsymbol{\Psi}|\boldsymbol{\Omega}, . &\sim IW(\delta + k - 1 + n, \boldsymbol{S}_b + \boldsymbol{\Omega}) \\
\boldsymbol{\Omega}|\boldsymbol{\Psi}, . &\sim W(\nu + \delta + k - 1, (\boldsymbol{\Psi}^{-1} + \boldsymbol{B}^{-1})^{-1})
\end{aligned}
$$

  with parameters as defined above. Given these posteriors, the Gibbs sampler implementation is straightforward.

Notation Conventions

- n number of clusters; i specific cluster
- J number of observations within cluster; j specific observation
- N total number of observations